BRAT: A Random Walk through the Semantic Spaces of the Blogosphere

Adil El Ghali, Yann Vigile Hoareau
Team Shakwat
CHArt – Lutin – Université Paris 8 *
2 rue de la Liberté – 93200 Saint Denis
{elghali, hoareau}@lutin-userlab.fr

1 Introduction

Semantic spaces, such as the *Latent Semantic Analysis* (LSA), *Hyperspace Analog to Language* (HAL) or *Random Indexing* (RI), offer convenient methods to represent semantic relations between words and concepts, abstracted from a distribution of documents. The distribution of documents determines the local co-occurrence pattern between words all over the corpus and, then, determines the semantic abstracted from the local distribution. Such methods are sensitive to the statistical properties on the distribution of words over documents. For instance, the semantic on the word *table* abstracted from a scientific corpus or a general corpus may be different. In the first case, since *table* may occur in the context of *table of correlation* or *table of results*, it would be considered to be associated to the word *correlation* whereas in the second case, because it may co-occur with *kitchen* or *living-room*, it would rather be considered as similar to *chair*. Nevertheless, the formal relation bearing the properties of the distribution of word's co-occurence and the final semantic produced by Semantic space methods have not been described until now. In the case of a mixed "scientific and general" corpus, what makes that the semantic of *table* became more similar to *chair* than *Speerman* and *vice-versa*?

We approached the Top-stories task of the Blog-Track'09 using a system named *Blogosphere Random Analysis using Texts* (BRAT) composed of two layers. The first layer distributes and represents blogs posts' in different semantic spaces built using Random Indexing. The second layer is an algorithm of retrieval that have the aim of navigate in the semantic space via a ramdom walk. BRAT have been constructed under two main working hypothesis that we considered important for dealing with the semantic of the blogosphere: the notion

*We are very grateful to the members of the DOXA project, for their material support, especially to Thalès and Pertimm

Report Documentation Page

Form Approved
OMB No. 0704-0188

Public reporting burden for the collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to a penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
NOV 2009		**00-00-2009 to 00-00-2009**

4. TITLE AND SUBTITLE	5a. CONTRACT NUMBER
BRAT: A RandomWalk through the Semantic Spaces of the Blogosphere	5b. GRANT NUMBER
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S)	5d. PROJECT NUMBER
	5e. TASK NUMBER
	5f. WORK UNIT NUMBER
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) **CHArt ? Lutin ? Universite Paris 8,2 rue de la Liberte,93200 Saint Denis,**	8. PERFORMING ORGANIZATION REPORT NUMBER
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	10. SPONSOR/MONITOR'S ACRONYM(S)
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES
Proceedings of the Eighteenth Text REtrieval Conference (TREC 2009) held in Gaithersburg, Maryland, November 17-20, 2009. The conference was co-sponsored by the National Institute of Standards and Technology (NIST) the Defense Advanced Research Projects Agency (DARPA) and the Advanced Research and Development Activity (ARDA).

14. ABSTRACT

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			
unclassified	**unclassified**	**unclassified**	**Same as Report (SAR)**	**10**	

of *semantic identity* and the notion of *semantic pollution*.

The article is organized as follows. In a first part, we shortly overview the methods and properties of semantic spaces models. The notions of semantic identity and semantic pollution are described in general together with their practical implication within the Top-stories task. In the second part, the BRAT system is described. The third part gives an overview of the performances of BRAT for the Top-stories task.

2 The cognition of Blog Mining

2.1 Semantic Spaces

Word Vectors are a family of models that represent semantic similarity between words in function of the textual environment in which those words appear. The words co-occurence distribution is collected, analyzed and transformed into a semantic space, in which words or concepts are represented as vectors in a high-dimension vector space. LSA [Landauer and Dumais, 1997], Hyper Analog to Language [Lund and Burgess, 1996] and Random Indexing [Kanerva et al., 2000] are some exemplars of Word Vectors. Those models are based on the Harris [Harris, 1968] distributional hypothesis, which states that words that appear in similar context have similar meanings. The definition of the unit of context is a common issue to all of those models, even if it is of different nature depending of the models. For example, LSA build a word-document matrix, in which each cell a_{ij} holds the frequency of a specific word i in a specific unit of context j. HAL defines a floating window of n words that scrolls each word of the corpus. Then build a word-word matrix, in which each cell a_{ij} contains the frequency a word i co-occurs with a word j for the considerate floating window. Different mathematical/statistical methods to abstract the meaning of concepts are applied on the distribution of frequencies stored in the word-document or word-word matrix. The first purpose of those mathematical processing is to abstract the central tendency of frequencies variations and to eliminating what can be considerate like "noise" caused by the part of specific use of language associated to each person or author. LSA uses a general method of linear decomposition of a matrix into principal independent components, which is called the Singular Value Decomposition (SVD). HAL reduces the expense of computational complexity by retaining a small number of principal components of the co-occurence matrix. Vectorial representations are used for the storage and the manipulation of concepts meaning. At the end of the process, similarity between two words may be calculated using different methods. A classical method is to calculate the value of the cosine of the angle between two vectors corresponding to a words or a group of words to approximate their semantic similarity. Another equivalent method is the pondered Euclidian distance.

In sum, Word Vectors inputs are a distribution of textual episodes defined as unit of context. The distribution of words co-occurence is matched with the

distribution of textual episodes in which they appear. Word Vectors outputs are concepts that emerged from this distribution's matching.

2.2 Semantic identity and pollution

As we started to introduce above, within the frame of semantic space methods, the semantic produced for a given word depends of the distribution of the other words that co-occur with it. It makes that no semantic of any words is given *ex nihilo, ie* pre-existing without (i) a learning process realized on (ii) a distribution of contexts or episodes (*ie*, unit of experience). The final semantic associated to a word have an identity that have been forged along the process of learning that is realized by SVD for LSA or the *accumulation* for RI. The semantic identity for a given word such as *table* changes in function of the corpus in appears within.

The notion of semantic identity addresses not only the scale of words but also the scale of the semantic space it-self. A semantic space have a particular identity that is given by the distribution of word's co-occurence that the space is composed by. The notion of semantic identity is circular because it reflects the circularity of the distributional hypothesis, which semantic spaces are based upon.

The notion of semantic identity does not produce something very new for researchers familiar with semantic spaces and the notion may appear somehow trivial if it did not allow to highlight a second notion that we will call the *semantic pollution*. In the previous example of a mixed "scientific and general" semantic space, the semantic identity of *table* is as much forged by the semantic related to science as by the semantic related to everyday life. In a general semantic space, if a word is similar to *table*, one can make the reasonable assumption that this word is not so far similar to *kitchen* or *house*. In a mixed "scientific and general" space, such an assumption became not so much reasonable, because the semantic of *table* have been some kind of polluted by the scientific part of the corpus. One can argue that this semantic pollution is nothing more than polysemy. It is true for the case of the word *table* because it is a polysemous word, but the pollution of the identity of the word *table* have and effect of pollution of the identity of words that it have co-occur together such as *correlation*, *Speerman*, *kitchen*, *house*, *etc*. Those words are not polysemous words but their semantic identity would be polluted too. Because of *table*, words such as *correlation* may possibly be not so far from *living-room* in term of semantic similarity. One again the semantic pollution addresses to the scale of word but also to the scale of the space for the same reason of circularity described above.

3 Application to Blogs Mining

The notion of semantic identity and semantic pollution are the two main ideas that are underlying our approach of the analysis of the blogosphere. In our view, the blogosphere is a cognitive system that produces textual information

that expresses people's views and ideas concerning views and ideas of others. For the Top-stories task of the Blog-Track of the TREC'09, the goals were (i) to detect the headlines of the New York Times that had produced exchanges in the blogosphere and (ii) for each of these headline, to propose some related blogs.

Considering the notion of semantic identity, we assume that the events of a given actuality produce some specific exchanges that are different of the exchanges produced relatively to the events of another actuality. Therefore, there is an advantage in splitting exchange in period of time in the aim of extracting the semantic identity associated with the actuality that have produce those exchanges.

Within the frame of semantic space models, the textual exchanges that are produced during a specific period of time constructs a semantic identity that is related to this particular actuality. Hence, in the first part of the process, semantic spaces are build from posts and commentaries written in a given period of time.

Nevertheless, even in choosing documents that have been produced during the same period of time, there is a large part of the selected exchanges that are not related to the headline of the New York Time. Those "not related texts" participated in the construction of the semantic identity corresponding to each semantic space, but they also pollute these semantic in the manner described above. The retrieval algorithm tries to navigate in a semantic space taking into account the degree of semantic pollution in the space.

4 BRAT

The Basic idea behind our work is that if we provide any efficient and easy way to navigate in a semantic space containing both blogs posts and headlines, then we can retrieve for each headline the relevant blogs posts by *walking randomly* in the semantic space. However we have to cope with the semantic pollution of the space.

The principle underlying the algorithm is to consider a representation of the "semantic identity of the day" as the sum of all document vectors corresponding to the given day. Taking into account that this representation might be strongly affected by a large amount of irrelevant documents, from the perspective of the top stories of the day. We defined a procedure that computes each document's similarity with both the "semantic identity of the day" and each of the headlines. In addition, for each headline, we rank a number of posts using a random walk through the semantic space. The procedure is stopped when satisfying a set of conditions that will be developed beyond.

Practically, for each topic and after a pre-processing phase, Random indexing [Sahlgren, 2006] was used to built a semantic space containing the blog posts, as well as the headlines, in a window around the date of the topic. This geometric representation of meanings of the episodes (posts and headlines) is then crawled using a random-walk-like algorithm to find the closest posts for

each headline. The ranking of the headlines takes into account the number of steps needed to find n relevant posts for a headline, together with the density of posts around the headline, as well as the average similarity between each headline and its associated posts. For each headline, the posts are ranked with regard to their similarity with the headline. Let us describe these steps with some more details.

4.1 The Blog08 data pre-processing

The Blog08 collection was made by crawling the blogosphere during more than year. The data were provided as-is: without any cleaning and the content of the blogs posts' where stored in a pseudo-XML format[1] which is unfortunetly not very well suited to store blogs data. The first not-very-interesting-but-necessary step was to split the permalinks files and organize them by posting date (instead of crawling date).

We also took the opportunity during this step to clean the posts from the parts that we consider useless such us: CSS and Javascript. but also to extract some general meta-data about posts and some structure informations of the blogosphere such us the *inter-comments network*.

The last step of the preparation of the data was to detect the languages of the blogs, in order to keep only english blogs. We use a language categorization library[2] that implements the algorithms described in [Cavnar and Trenkle, 1994] to categorize texts using n-grams.

4.2 Semantic space construction

The Semantic space method we use in the context of the Blog-Track'09 is Random Indexing (RI), which is not a typical method in the family of Semantic space methods. Particularities of RI are that (i) it does not create co-occurrence matrix (but it is possible if needed) and (ii) it does not need heavy statistical treatments like SVD for LSA. Contrary to the other Word Vector models, RI is based on random projection, a method that approximate statistics co-occurences, and allows to scale to huge number of documents. The construction of a semantic space with RI is as follows:

- Create a matrix $A(d \text{ x } N)$, containing Index vectors, where d is the number of documents or contexts and N, the number of dimensions ($N > 1000$) decided by the experimenter. Index vectors are sparse and randomly generated. They consist in small numbers +1 and -1 and thousands of 0.
- Create a matrix $B(t \text{ x } N)$, containing term vectors, where t is the number of different terms in the corpus. Set all vectors with null values to start the semantic space construction.

[1] The files of the Blog08 collection are not in a well formed XML format, and the preparation of the data was a very time and resource consuming task

[2] http://olivo.net/software/lc4j/

- Scan each document of the corpus. Each time a term τ appears in a document δ, accumulate the randomly generated δ-index vector to the τ-term vector.

At the end of the process, term vectors that appeared in similar contexts have accumulated similar index vectors. There is a training cycle option in the model. When the scan has been computed for all documents, the matrix B is charged for all term vectors. Then a matrix $A'(d'$ x $N)$, with d' = d can be computed with the output of term vectors. The number of training cycle is a parameter in the model. The training process improves the quality of the Semantic space. The RI model has performed in TOEFL synonymy test [Kanerva et al., 2000, Karlgren and Sahlgren, 2001] as well as in text categorization [Sahlgren and Cöster, 2004].

For each topic (a date D) a semantic space SS_D is built relying on the Semantic Vectors[3] library [Widdows and Ferraro, 2008]. The semantic space contains two kinds of episodic documents: (i) all the headlines in a window[4] [$D-1$, $D+1$], (ii) all the english posts[5] in a window [$D-1$, $D+3$].

4.3 A random walk in the semantic space

Once the semantic space SS_D of a day D constructed, we use a random-walk-like algorithm to navigate in the space in order to retrieve for each headline n related blog posts.

We call a prototype for a category of a set of documents (blog posts or headlines), a pseudo document represented in the semantic space by the sum of all the vectors in the set. For instance, the prototype of all the headlines is a pseudo document P_H represented by the vector:

$$\vec{P_H} = \sum_{h \in H} \vec{h} \tag{1}$$

where H is the set containing all the headlines of SS_D.

Given a headline $h_i \in SS_D$ and $\eta \in \mathbb{N}$, we call $\eta-$neighbourhood of h_i w.r.t a prototype P, the set of blogs posts defined as follow:

$$\eta - \text{neighbourhood}(h_i, P) = \{b_j | d(b_j, h_i) < \frac{d(P, h_i)}{\eta}\} \tag{2}$$

where $d(d_i, d_j)$ is an euclidien distance in the semantic space between the vectors $\vec{d_i}$ and $\vec{d_j}$.

In order to retrieve the n related blog posts for the headline h_i, we choose a threshold $m > n$, we walk randomly through the set B containing all the blog posts of SS_D until founding m candidates posts in the $\eta-$neighbourhood of h_i w.r.t the prototype P_H of all the headlines. If we found m candidates posts, we

[3] http://code.google.com/p/semanticvectors/

[4] Except for the run ri2049rw3 where only the headlines of the day D were considered.

[5] We choose to consider as an episode the document containing a blog post and its comments.

define the score p_i of the headline h_i as the number of steps we walked in B. If the number of founded blog posts $m' < m$ then the score p_i of h_i is defined as

$$p_i = card(B) - m' \tag{3}$$

In each set B_i containing the founded blog posts for h_i, we keep the $min(n, m')$ closest blog posts to h_i as the related posts. And the headlines are ranked in ascending order of p_i.

5 Results

The submitted runs implement different hypothesis concerning the organisation of the knowledge in semantic space build from the blogosphere. The runs correspond to different values of η used for the random walk algoritm.

In the runs ri2049rw3 corresponds to an application of the algorithm with $\eta = 3$ in a 2049 dimensions space.

The run ri1025rw5432 corresponds to an adaptative algorithm using the same principle, and where the results of the random walk with $\eta = 5, 4, 3, 2$ are combined.

The run ri1025rw5h2b uses a similar algorithm but with little modification of the definition neighbourhood. The used neighbourhood is the intersection of the 5-neighbourhood w.r.t to P_H and the 2-neighbourhood w.r.t to P_B.

The run ri1025rw2b corresponds to an application of the algorithm with the 2-neighbourhood w.r.t to P_B.

Run ID	R-Precision $\geq$ Median	% of Retrieved Relevant Headlines
ri1025rw2b	32	24%
ri1025rw5432	32	24%
ri1025rw5h2b	34	24%
ri2049rw3	26	20%

Table 1: Comparaison of R-Precision and Number of Relevant Retrieved Headlines with the Median values

The obtained results are summerized in Table 1 and Figure 1. The good performance of the *adaptative* run (ri1025rw5432) and the even better performance of the *double constraint* run (ri1025rw5h2b) constitutes good arguments in favour of the validity of the notion of semantic identity and semantic pollution.

6 Conclusions

The original contribution of our work is to propose a simple and efficient algorithm to navigate in a semantic space in which the semantic of blog posts is supposed to be strongly polluted because of a number of irrelevant posts. The

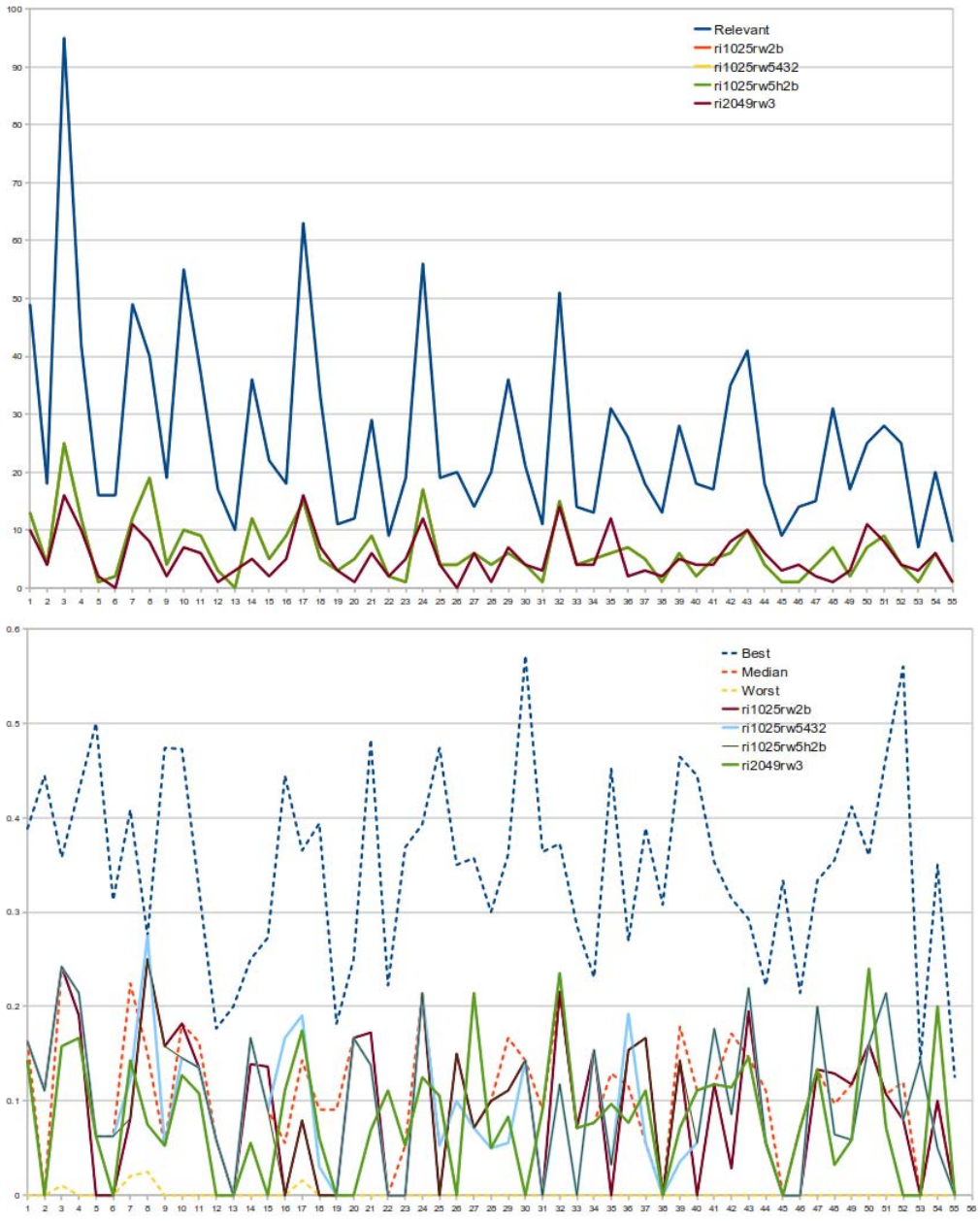

Figure 1: Retrieved relevant posts and R-Precision for the 4 runs with reference values

principle underlying the algorithm is to consider a representation of the "semantic identity of the day" as the sum of all document vectors corresponding to the given day. Taking into account that this representation might be strongly affected by a large amount of irrelevant documents, from the perspective of the top stories of the day. We defined a procedure that computes each document's similarity with both the "semantic identity of the day" and each of the headlines. In addition, for each headline, we rank a number of posts using a random walk through the semantic space.

Acknowledgment

This work would have never been possible without the support of Charles Tijus and the Lutin Lab. We especially want to thank the Lutin members Zakia Ikhlef, Rebecca Djuric, Daniel Hromada, Olivier Floucat, and Françoise Richard for their valuable support.

We are also grateful to the members of the DOXA project and the Cap Digital Business Cluster, especially Thibaut Ehrette, Jacques Bibal, Catherine Gouttas from Thalès, Patrick Constant and Guillaume Logerot from Pertimm.

References

[Cavnar and Trenkle, 1994] Cavnar, W. B. and Trenkle, J. M. (1994). N-gram-based text categorization. In *In Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 161–175.

[Harris, 1968] Harris, Z. (1968). *Mathematical Structures of Language*. John Wiley and Son, New York.

[Kanerva et al., 2000] Kanerva, P., Kristoferson, J., and Holst, A. (2000). Random Indexing of Text Samples for Latent Semantic Analysis. In Gleitman, L. and Josh, A., editors, *Proceedings of the 22nd Annual Conference of the Cognitive Science Society*, Mahwah. Lawrence Erlbaum Associates.

[Karlgren and Sahlgren, 2001] Karlgren, J. and Sahlgren, M. (2001). From Words to Understanding. In Uesaka, Y., Kanerva, P., and Asoh, H., editors, *Foundations of Real-World Intelligence*. CSLI Publications, Stanford.

[Landauer and Dumais, 1997] Landauer, T. K. and Dumais, S. T. (1997). A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge. *Psychological Review*, 104(2):211–240.

[Lund and Burgess, 1996] Lund, K. and Burgess, C. (1996). Producing high-dimensional semantic space from lexical co-occurence. *Behavior research methods, instruments & computers*, 28(2):203–208.

[Sahlgren, 2006] Sahlgren, M. (2006). *The Word-Space Model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces*. PhD thesis, Department of Linguistics Stockholm University.

[Sahlgren and Cöster, 2004] Sahlgren, M. and Cöster, R. (2004). Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 487, Morristown, NJ, USA. Association for Computational Linguistics.

[Widdows and Ferraro, 2008] Widdows, D. and Ferraro, K. (2008). Semantic Vectors: A Scalable Open Source Package and Online Technology Management Application. In *Proceeding of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*.